Pika, Anastasiia, van der Aalst, Wil M.P., Fidge, Colin J., ter Hofstede, Arthur H.M., & Wynn, Moe T. (2012) Predicting deadline tansgressions using event logs. In *Lecture Notes in Business Information Processing*, Springer, Tallin, Estonia, pp. 211-216.

http://dx.doi.org/10.1007/978-3-642-36285-9_22

Predicting Deadline Transgressions Using Event Logs

Anastasiia Pika[1], Wil M. P. van der Aalst[2,1], Colin J. Fidge[1],
Arthur H. M. ter Hofstede[1,2], and Moe T. Wynn[1]

[1] Queensland University of Technology, Brisbane, Australia
{a.pika,c.fidge,a.terhofstede,m.wynn}@qut.edu.au
[2] Eindhoven University of Technology, Eindhoven, The Netherlands
{w.m.p.v.d.aalst}@tue.nl

Abstract. Effective risk management is crucial for any organisation. One of its key steps is risk identification, but few tools exist to support this process. Here we present a method for the automatic discovery of a particular type of process-related risk, the danger of deadline transgressions or overruns, based on the analysis of event logs. We define a set of time-related process risk indicators, i.e., patterns observable in event logs that highlight the likelihood of an overrun, and then show how instances of these patterns can be identified automatically using statistical principles. To demonstrate its feasibility, the approach has been implemented as a plug-in module to the process mining framework ProM and tested using an event log from a Dutch financial institution.

1 Introduction

Effective risk management is crucial for organisations. ISO Guide 73:2009 defines risk as the "effect of uncertainty on objectives" where effect is "a deviation from the expected — positive and/or negative" [3]. One of the most important aspects of risk management is *risk identification* [7]. Traditional risk management approaches offer only high-level guidance about risk identification methods and rely on the knowledge of domain experts [7]. Accordingly, our goal is to show how the data recorded in event logs by contemporary workflow management systems can be exploited for the purpose of risk identification.

Various approaches for predicting timeliness have been proposed in the literature [8, 9] and serve as a starting point for our work. Van der Aalst et al.'s approach [8] builds an annotated transition system and remaining process time is then predicted based on the average of earlier cases visiting the same state. Van Dongen et al.'s approach [9] predicts the remaining cycle time of a case by using non-parametric regression based on case-related data as the predictor variables. A framework for identification and analysis of the operational risks associated with single business process activities, as well as a whole process, was proposed by Jallow et al. [4]. Wickboldt et al. proposed a framework that makes use of a process model and process execution data from historical records for risk prediction [10]. The use of process mining for the identification of transactional fraud risk was proposed by Jans et al. [5]. Overall, our approach differs from previous work in that: it does not require as an input risk indicators defined by experts or pre-classified data [4, 10]; it is not restricted to transactional fraud risk [5]; and it focuses on identifying the risk of not meeting a deadline rather than estimating the remaining cycle time of a case [8, 9].

Since our approach is based on actual data in event logs, it focuses on *process-related* risks only. We refer to a risk as *process-related* if its root cause is any combination of process behaviour (notably the activities performed and their sequence), resource behaviour (e.g., resource availability, capabilities and interaction patterns) or case-related data. Process-related risks can jeopardise the achievement of process goals in terms of cost, timeliness or the quality of outputs [4]. In this paper we consider only one type of risk, the likelihood that cases do not meet their deadline, however our general strategy is not restricted to time-related risks. Our approach consists of three steps: 1) definition of Process Risk Indicators (PRIs); 2) devising a way to identify instances of risk patterns in a log; and 3) defining a predictor function that characterises the risk of a case failing (from its local characteristics only).

2 Risk Identification Method

Before introducing our Process Risk Indicators (PRIs), we first introduce some notations. Let α denote a *run* of a *process model*. Random variable X_α denotes a case's outcome in terms of timeliness per run α. In this paper, we assume that X_α takes one of two possible values: *1* if a case is delayed and *0* if it is completed in time. Per run α there is *cumulative distribution function* F_α such that $F_\alpha(x) = P(X_\alpha \leq x)$ for X_α. In this way the risk of case delay can be quantified. Function F_α captures both impact and likelihood. Assuming that a process is in a steady state there exists such a function F_α for all runs. Our goal is to define a function G that predicts the value of X_α, i.e., we would like to minimize the expected value of the difference $|X_\alpha - G_\alpha|$. Function G_α is based on a few local characteristics of α. Let $\mathcal{E}$ denote the set of all possible *events*. A *trace* is a sequence of events $\delta \in \mathcal{E}^*$. An *event log* L is a set of traces. We assume that each event has the following attributes: an *activity name*, a *time stamp*, a *resource* and a *transaction type* (including *start* and *complete*). Each case is described by a *trace* $\delta \in L$ which can be related to a process model *run*.

Using indicators for risk monitoring is a common practice in areas such as safety and fraud detection, so we use "risk indicators" for identification of process-related risks. We define a *Process Risk Indicator* as a pattern observable in an event log whose presence indicates a higher likelihood of some process-related risk. In this paper we consider only the risk of a case overrun. Our aim is to identify domain-independent indicators that can be identified by analysing event logs and do not require any additional information, e.g. a process model. We have defined five time-related PRIs.

– **PRI 1: Abnormal activity execution time.** A case contains an activity whose duration is significantly higher than its normal duration.

– **PRI 2: Abnormal waiting time.** Activity execution is not started for an abnormally long period of time after it has been enabled.
– **PRI 3: Multiple activity repetitions.** An activity is repeated multiple times in a case.
– **PRI 4: Atypical activities.** A case contains an activity that has not been performed often previously.
– **PRI 5: Multiple resource involvement.** The number of resources involved in a case significantly exceeds the norm.

Our method for PRI discovery is based on unsupervised statistical techniques for outlier identification. They have the advantage of not requiring pre-classified data samples for learning. We use the "sample standard deviations" approach for outlier detection which assumes that the sampled values follow a normal distribution. A cut-off threshold for a normally distributed population is usually defined as $\mu \pm 2\sigma$ (for a 95% confidence interval). Observations whose values are outside this range are considered outliers. If a sample contains extreme outliers a cut-off threshold defined by the mean $\overline{x}$ and standard deviation s is often unnecessarily biased, so for a normally distributed population the median $\tilde{x}$ is a robust estimator for $\overline{x}$ and a robust estimator for s is 1.483MAD [6]. Our method for PRI identification consists of two steps: (1) Identify a cut-off threshold by analysing the given event log; and (2) For a given case (represented by a trace) identify outliers using the learned threshold. For each trace $\delta \in L$ we introduce attributes for each risk indicator n, denoted PRI_n. These attributes are used by the risk identification method to store information about the indicators found in a trace. Attribute PRI_n is 1 if indicator n is found, and 0 otherwise.

Following Zhang et al. [11], we assume that activity durations follow a log-normal distribution, therefore logarithms of activity durations approximately follow a normal distribution. To identify the presence of PRI_1 in a trace belonging to a run α of the process model, the following procedure is followed. For every activity a occurring in at least one trace corresponding to α: create a sample x of logarithms of the durations of all occurrences of a in traces corresponding to α (difference between *complete* and *start* events); calculate a cut-off threshold $t = \overline{x} + 2s$; for a given activity instance compare logarithm of its duration with the threshold t and if it exceeds the threshold set the value of the corresponding case's attribute $PRI_1 = 1$. A similar procedure is followed for other PRIs. For PRI_2 we also assume that waiting times follow a log-normal distribution [11]. The waiting time is calculated as the difference between the end time and the start time of two consecutive activities in a log. Importantly, this assumption may not always be true. For PRI_3 and PRI_5 we assume that the number of activity executions in a case and the number of resources involved in a case follow a normal distribution. An activity is considered atypical (PRI_4) if it has been executed in fewer than a certain number of cases in the log. The threshold t is an input parameter that represents the fraction of cases where a particular activity has been executed.

We define a predictor function G that estimates the risk level of a case based on the risk indicators it exhibits. Thus binary function G predicts a delay if *any* of the indicators is found in a case. We have also defined a function $Score$ that returns a "suspicion score" based on the number of identified indicators for each case. A high suspicion score means that many indicators were found in a case, and can be used to calibrate risk alert levels. Let δ be a trace that represents a given case, $\delta(PRI_n)$ denote the value of attribute PRI_n of trace δ, $\{PRI_1, \dots, PRI_k\}$ be a set of k PRIs, and w_i denote the weight chosen for indicator PRI_i:

$$G(\delta) = \bigvee_{i=1}^{k} \delta(PRI_i);\ Score(\delta) = \sum_{i=1}^{k} w_i * \delta(PRI_i)$$

In our current implementation once a risk indicator is identified we update the corresponding attribute of a trace. Functions G and $Score$ are calculated for each complete trace and the values are compared with actual case durations to evaluate the performance of the functions.

3 Experimental Results

Our approach has been implemented as a plug-in of the process mining framework ProM 6. Its main functionality is to identify occurrences of our five PRIs in a given log and to thus predict the likelihood of a case being delayed. Predicted values are then compared with the actual outcome of a case to evaluate the performance of the predictor functions. In order to isolate traces corresponding to different process model runs the plug-in uses either the existing ProM 6 "replay" plug-in [1] or the trace clustering plug-in [2] (if the process model is not available). We evaluated our approach using an event log which represents the application process for a personal loan or overdraft from a Dutch financial institution given for the BPI Challenge 2012.[3] The log contains 13,087 traces in total and we first filtered this log to produce 934 traces suited to our experimental purposes. The plug-in that uses the trace clustering was applied. The filtered log was grouped into 12 clusters with the total number of traces in each cluster ranging from 20 to 206. After clustering, the traces in each cluster were put into either a training set (used to learn cut-off thresholds) or a test set. For each cluster within the training set we estimated the normal case duration as $\tilde{x} + 2 * 1.483 * MAD$. Cases whose durations exceeded this value were considered to be delayed.

Table 1 shows the experimental results for the test set of 462 traces. To evaluate the quality of predictions we used the mean absolute error (MAE). This is calculated as $\frac{1}{n}\sum_{i=1}^{n}|p_i - r_i|$ for both delayed cases (yielding the MAE for false negatives) and for cases that are in time (yielding the MAE for false positives), where n is the number of cases in each category and p_i and r_i denote predicted and real values respectively. We calculated the MAE separately for delayed cases and cases that are on time, because it is often important to distinguish between different types of errors, both false-negatives and false-positives, as their impact

[3] BPI Challenge 2012. doi:10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f

Table 1. Experimental results showcasing the predictive value of five process risk indicators (PRIs) on the test set of the BPI Challenge event log.

	5 PRIs				PRI 1		PRI 2		PRI 3		PRI 4		PRI 5	
	Delayed		In Time											
	TN	FN	FP	TP	TN	FP	TN	FP	TN	FP	TN	FP	TN	FP
Traces	22	7	221	212	8	115	0	19	19	121	0	2	3	25
%	76%	24%	51%	49%	28%	27%	0%	4%	66%	28%	0%	0.5%	10%	6%

Legend: TN—True Negatives; FN—False Negatives; FP—False Positives; TP—True Positives

on business performance can be very different. We can observe that the MAE for delayed cases with 5 PRIs is 0.24, i.e., the predictor function estimated correctly the outcome of 76% of delayed cases ("True Negatives" in Table 1). On the other hand, the MAE for the cases that are not delayed is 0.51 ("False Positives" in Table 1). From further analysis, we observed that 74% of the 221 cases that were falsely predicted as delayed have durations that are very close to the cut-off threshold (the difference is lower than 5% of assumed normal case duration). From the individual PRI results, we can see that for this particular log almost all predicted problems ("True Negatives" in Table 1) are based on observations of PRIs 1, 3 and 5. We have also analysed the ability of PRIs to provide operational support. For this particular event log, we were able to identify the presence of PRIs 1, 3 and 4 early during a case's execution, while PRIs 2 and 5 for most of the cases could only be discovered after half of the normal case duration for the run corresponding to that case had passed.

Table 1 focussed on the results from our first predictor function, G. We also tested the weighted *Score* function (with $w_i = 1$ for all PRIs) and found that for most of the cases predicted as delayed just one of the indicators was discovered (64% of correctly predicted cases and 76% of falsely predicted cases). This reveals that the "suspicion" attached to these poor results of G was actually very low.

After examining the BPI Challenge event log we noted certain log characteristics that may have influenced the presented results and discovered opportunities for the improvement of the risk identification method. The durations of the cases assigned to a cluster did not significantly deviate from the cut-off thresholds, thus there were very few outlier cases. Also, the number of traces in some clusters were too small to get statistically significant results. Many activities have very small durations compared to the total case duration. Discarding durations whose values are lower than some predefined threshold may help to filter out false positive predictions. The event log used does not contain *start* events recorded for all activities. To be able to work with the event logs that do not contain *start* events we can use an indicator "PRI 6: Abnormal sub-process duration" that considers both activity service and waiting time (sub-process durations are calculated as the time difference between two consecutive *complete* events). Applying PRI 6 and PRI 3 v.2 (that considers the absolute values of repetition durations) we were able to correctly estimate the outcome of 86% of delayed cases and 30% of cases in time were falsely predicted as delayed.

4 Conclusions

We have presented a new approach for predicting whether or not a case will meet its deadline. We first defined relevant "Process Risk Indicators" and then used statistical methods to identify their presence in event logs. Our initial results indicate that further work is needed to properly calibrate the analysis, perhaps on a process-specific basis, to minimise the annoyance of false-positive warnings and the more serious threat of false-negative alert failures. (As noted above, the data set available to us for experimentation was not well-suited to our purposes. We have recently obtained a larger data set from an Insurance Company and will use it for experiments.) Although we only focused on the risk of case overruns in this paper, we believe that the overall strategy is suitable for any quantifiable type of risk, such as financial losses or low-quality outputs.

Acknowledgement. This research is funded by the ARC Discovery Project "Risk-aware Business Process Management" (DP110100091).

References

1. A. Adriansyah, B.F. van Dongen, and W.M.P. van der Aalst. Conformance checking using cost-based fitness analysis. In *Enterprise Distributed Object Computing Conference (EDOC), 2011 15th IEEE International*, pages 55–64. IEEE, 2011.
2. R. Bose and W.M.P. van der Aalst. Context aware trace clustering: Towards improving process mining results. In *Proceedings of the SIAM International Conference on Data Mining, SDM*, pages 401–412, 2009.
3. International Organization for Standardization. *Risk management: vocabulary = Management du risque: vocabulaire (ISO guide 73).* Geneva, 2009.
4. A.K. Jallow, B. Majeed, K. Vergidis, A. Tiwari, and R. Roy. Operational risk analysis in business processes. *BT Technology Journal*, 25(1):168–177, 2007.
5. M. Jans, N. Lybaert, K. Vanhoof, and J.M. van der Werf. A business process mining application for internal transaction fraud mitigation. *Expert Systems with Applications*, 38(10):1335113359, 2011.
6. P.J. Rousseeuw. Robust estimation and identifying outliers. In *Handbook of Statistical methods for Engineers and Scientists*, chapter 16. McGraw-Hill, New York, 1990.
7. Standards Australia and Standards New Zealand. *Risk management: principles and guidelines (AS/NZS ISO 31000:2009).* Sydney, NSW, Wellington, NZ, 3rd edition, 2009.
8. W.M.P. van der Aalst, M.H. Schonenberg, and M. Song. Time prediction based on process mining. *Information Systems*, 36(2):450–475, 2011.
9. B. van Dongen, R. Crooy, and W.M.P. van der Aalst. Cycle time prediction: When will this case finally be finished? *On the Move to Meaningful Internet Systems: OTM 2008*, pages 319–336, 2008.
10. J.A. Wickboldt, L.A. Bianchin, R.C. Lunardi, L.Z. Granville, L.P. Gaspary, and C. Bartolini. A framework for risk assessment based on analysis of historical information of workflow execution in it systems. *Computer Networks*, 55(13):2954–2975, 2011.
11. P. Zhang and N. Serban. Discovery, visualization and performance analysis of enterprise workflow. *Computational statistics & data analysis*, 51(5):2670–2687, 2007.